

MAXIMUM LIKELIHOOD ESTIMATES OF PAIRWISE REARRANGEMENT DISTANCES

STUART SERDOZ¹, ATTILA EGRI-NAGY¹, JEREMY SUMNER², BARBARA R. HOLLAND², PETER D. JARVIS², MARK M. TANAKA³, AND ANDREW R. FRANCIS¹

ABSTRACT. Accurate estimation of evolutionary distances between taxa is important for many phylogenetic reconstruction methods. In the case of bacteria, distances can be estimated using a range of different evolutionary models, from single nucleotide polymorphisms to large-scale genome rearrangements. In the case of sequence evolution models (such as the Jukes-Cantor model and associated metric) have been used to correct pairwise distances. Similar correction methods for genome rearrangement processes are required to improve inference. Current attempts at correction fall into 3 categories: Empirical computational studies, Bayesian/MCMC approaches, and combinatorial approaches. Here we introduce a maximum likelihood estimator for the inversion distance between a pair of genomes, using the group-theoretic approach to modelling inversions introduced recently. This MLE functions as a corrected distance: in particular, we show that because of the way sequences of inversions interact with each other, it is quite possible for minimal distance and MLE distance to differently order the distances of two genomes from a third. This has obvious implications for the use of minimal distance in phylogeny reconstruction. The work also tackles the above problem allowing free rotation of the genome. Generally a frame of reference is locked, and all computation made accordingly. This work incorporates the action of the dihedral group so that distance estimates are free from any *a priori* frame of reference.

1. INTRODUCTION

Estimates of evolutionary distance between pairs of taxa are key ingredients for reconstructing phylogenies, but are difficult to obtain reliably—see [Felsenstein \[2004\]](#), [Gascuel \[2005\]](#). This is especially true for evolutionary models in which events can interact with each other in a way that affects inference, such as most rearrangement models in bacteria. One estimate of distance between two genomes is the *minimal* distance which is model-specific and represents an assumption of parsimony in evolutionary paths through genome space (see [Fertin \[2009\]](#) for a discussion of rearrangement models in this context). In fact there are infinitely many possible evolutionary paths between any two genomes, and the minimal distance is simply the length of one of the shortest of these. Therefore the minimal distance often underestimates the true number of evolutionary events.

The problems with using a minimal distance are well documented, especially when time periods are long and the space of obtainable genomes becomes saturated. Given enough time, all evolutionary endpoints become equally likely, and any signal

¹CENTRE FOR RESEARCH IN MATHEMATICS, WESTERN SYDNEY UNIVERSITY

²SCHOOL OF PHYSICAL SCIENCES, UNIVERSITY OF TASMANIA

³SCHOOL OF BIOTECHNOLOGY AND BIOMOLECULAR SCIENCES, UNIVERSITY OF NEW SOUTH WALES

Date: July 11, 2016.

of actual evolutionary time is lost. In some models, metrics have been developed to account for multiple changes; the most well-known perhaps being the Jukes-Cantor correction for models of single nucleotide substitution [Jukes and Cantor, 1969]. This method requires all events to be *independent* (a common assumption with nucleotide substitution), but independence among sites does not hold for most genome rearrangement models (such as inversion) and so alternative approaches are needed.

Given pairwise distances that are obtained from a phylogenetic tree, Buneman [1971] demonstrated that the tree is uniquely recoverable from the distances, a fact which also follows from the 4-point condition [Buneman, 1974]. Furthermore, Warnow [1996] and Atteson [1999] suggest that if the true evolutionary distance inference is sufficiently accurate, even polynomial time reconstruction algorithms, such as Neighbor Joining [Saitou and Nei, 1987], will return the correct phylogeny. Recent work by Gascuel and Steel [2015] places the results of Atteson *et al.* in a statistical framework.

Some empirical studies attempt to find a relationship between true distance and minimal distance (or some other available measure such as breakpoint distance). The end product is an estimate of true distance as a function of minimal distance. For instance, Wang and Warnow [2001] introduced an estimator of true evolutionary distance called *IEBP* (inverting the expected breakpoint distance). The method operates under the generalised Nadeau-Taylor model [Nadeau and Taylor, 1984] and provides a robust polynomial time algorithm to estimate true evolutionary distance. Similarly, the *EDE* (empirically derived estimator) of Moret *et al.* [2001] samples the relationship between inversion distance and true evolutionary distance before providing a fit. Applications of IEBP and EDE can be seen in Li-San [2002].

While a useful correction, such estimates are based on just one factor – the minimal distance – and can’t take into account the underlying structure of the paths in genome space (in our framework, the Cayley graph of the group). The key point being, that not all elements of equal minimal distance are equally likely.

Given the sizes of the spaces involved, MCMC and Bayesian methods play an important role. York *et al.* [2002] uses a Bayesian framework to estimate true distances for inversions. On the MCMC front, Miklós [2003] introduced a time continuous stochastic approach to genome rearrangements (modelled as a Poisson process), allowing reliable estimates of true distances. The key aim being to describe the posterior distribution of true evolutionary distance given 2 arrangements. There have been several generalizations to these methods: Durrett *et al.* [2004] includes translocations as well as inversions. ? describe a Bayesian method for phylogeny inference and offer a comparison between thier approach and a parsimony approach. Miklós and Darling [2009] provide a method to estimate the *number* of minimal paths.

As an optimal estimate of true distance we would like (very loosely) some sort of *expected* distance constructed as a weighted average of evolutionary paths, pushing the problem into the intersection of combinatorics and statistics. In this vein, Eriksen [2002] offered an approximation of the expected number of inversions to have occurred given n breakpoints. This was followed by a method of estimating the expected inversion distance by looking at the expected transposition distance [Eriksen and Hultman, 2004]. Generalisations by Eriksen *et al.* include Eriksen [2005] and Dalevi and Eriksen [2008].

This paper describes a *maximum likelihood* approach to rearrangement distances, as an alternative to minimal distance. We focus on models of genome rearrangement involving invertible operations, such as inversion and translocation, which can be described in group-theoretic terms, using the framework introduced in Egri-Nagy et al. [2014b] and Francis [2014]. This algebraic framework treats genomes as the images of the actions of elements of a finite reflection group, and allows us to treat the genome as not fixed in space, but free to rotate in three dimensions. Each genome is then considered to be a coset in the quotient of the main reflection group by the dihedral group.

The maximum likelihood estimator (MLE) we present is built upon two key assumptions: firstly, rearrangement events occur according to a Poisson process; and secondly, sequences of evolutionary events of equal length are equally probable.

The next Section describes the general group-theoretic models of chromosome rearrangements on which this paper is based. Section 3 introduces the likelihood function under our model, and gives some basic examples of what these functions look like. Section 4 compares the minimal distance to the MLE and gives an example of how the resulting phylogenetic inference can give different results. Section 5 looks at potential properties of group elements which may characterise the likelihood function and hence the MLE of distance. Section 6 describes what is required to account for dihedral symmetry. We end with a discussion of some of the issues involved in using the MLE.

2. GROUP THEORETIC MODELS OF REARRANGEMENT

In this section we describe group-theoretic models of genome rearrangement, following the development in Egri-Nagy et al. [2014b]. Such models allow events that change the underlying sequence in a reversible way, including for example inversion and translocation but not insertion, excision, or horizontal gene transfer. The invertible rearrangements defined by the model then generate a *group*, and there is a one-to-one correspondence between the set of possible genome arrangements and the set of elements of this group.

This correspondence in practice requires two additional assumptions. First, we choose one genome as the reference genome, that will correspond to the group identity element. This is arbitrary, and is discussed in more detail below. Second, we assume there is no rotation of the genome in 3-dimensional space. We think of this as fixing a “frame of reference” for all genomes. This assumption is removed for calculating MLEs of evolutionary distances in ways described below, by taking a quotient by the dihedral group.

The genome space is then realized as a graph with genomes as vertices and allowable evolutionary events defining edges between them. This corresponds to a graph based on the group, called the *Cayley graph*, whose vertices are group elements and edges represent multiplication by the group generators. Thus the Cayley graph can be thought of as a map of the genome space, with vertices the possible genomes (group elements) and edges the possible rearrangement events (generators of the group) [Clark et al., 2016]. The Cayley graph depends on both the group \mathcal{G} and the generating set \mathcal{S} . For a general reference to random walks on Cayley graphs see Aldous and Fill [2002].

Given a choice of one arrangement as the reference genome G_0 , every other genome arrangement can be obtained from G_0 by a sequence of rearrangements. Because

each allowable rearrangement event defines a generator of the group, this sequence of rearrangements is a product of group generators, and therefore corresponds to a group element itself. Thus the reference genome G_0 corresponds to the identity element e of the group \mathcal{G} , and each other possible genome corresponds to a unique group element (remembering that for now we assume a fixed frame of reference). Note that there may be many sequences of events giving rise to the same genome, and these correspond to different paths through the Cayley graph.

The choice of reference genome is not important. For any two genomes G_1 and G_2 with corresponding group elements g_1 and g_2 , there is a unique group element (namely $g_1^{-1}g_2$ when acting on the right) that transforms G_1 into G_2 . As a result of the transitive group action [Babai, 1996], the group element is independent of the choice of reference genome. For instance if G_1 was chosen as the reference genome then the path from G_1 to G_2 would still correspond to the group element $g_1^{-1}g_2$.

With this correspondence between the genome space and the Cayley graph, the *minimal distance* (denoted d_{min}) on the genome space, the “word metric” on the group [Lyndon and Schupp, 1977], and the path metric on the Cayley graph all coincide.

An evolutionary history between two genomes is a random walk on the genome space using allowable rearrangements, or equivalently, a random walk on the Cayley graph — a well-studied topic [Aldous and Fill, 2002, Godsil and Royle, 2001, Lubotzky, 1995]. Such a walk corresponds to a sequence of (right) multiplications of the group element at the starting point by the generators labelling the edges on the path. That is, an evolutionary path from g_1 to g_2 takes the form of the initial genome followed by a concatenation on the right of the applied events. For example, a path along the edges from g_1 beginning with s_2 and subsequently the sequence of generators s_5, s_2, s_1, s_7 corresponds to the equation in the group given by $g_1 s_2 s_5 s_2 s_1 s_7 = g_2$. This corresponds to a walk or path of length 5. Each such product of group generators from a path between g_1 and g_2 represents the same group element, namely $g_1^{-1}g_2$.

The transitivity that we mentioned earlier means that paths from g_1 to g_2 are in correspondence with paths from the identity e to $g_1^{-1}g_2$, and so it is sufficient to study paths and distances from the identity to a group element g .

Typically there are many paths to a group element, each giving a distinct word in the generators of the group (labels on edges of the Cayley graph). A *reduced* word is one that corresponds to a minimal length path. Any word in the generators can be reduced to a minimal one using the group relations; these in turn correspond to loops in the Cayley graph. Generally reduced words are not unique (accounting for the prospect of multiple paths of minimal length). For further reading on the interaction between relations and words see Lyndon and Schupp [1977]. In what follows we will not be just interested in paths of minimal length, but in all paths between two genomes.

In this set-up, given a random walk from e to g the minimal distance is the length of a geodesic path, while the true evolutionary distance is the length of the actual path. The model of rearrangement we use as an example throughout this paper is the 2-inversion model studied by Egri-Nagy et al. [2014b], in which adjacent regions are swapped, and orientation is ignored. However our general principles apply to any group-theoretic rearrangement system in which the generators are of order 2, and in particular any model of inversions.

When referencing specific circular genome arrangements we will use cycle notation in which the cycle (\dots, a, b, c, \dots) means “ \dots , region a is in position b , and region b is in position c , \dots ”. For instance the permutation shown in Figure 1 is represented by $g = (3, 7, 5)(4, 6)$, which we read “region 3 is in position 7, region 7 is in position 5, region 5 is in position 3; region 4 is in position 6 and region 6 is in position 4”. Swapping regions 4 and 5 is done by multiplying on the right by the generator $(4, 5)$, which gives the result $(3, 7, 5)(4, 6)(4, 5) = (3, 7, 4, 6, 5)$ (the reader may draw this to convince herself that this has the desired result).

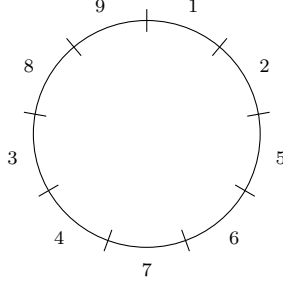


FIGURE 1. Cycle notation tracks the movement of individual regions. The genome above is represented by $g = (3, 7, 5)(4, 6)$.

2.1. Genomes in three dimensional space. Fixing the frame of reference makes for a clean translation of genome space to the Cayley graph, but in fact if two arrangements can be obtained from each other by rotation in three dimensional space, then biologically they are the same arrangement. Mathematically, this means that if two circular genomes can be reached from each other by rotating the circle or reflecting it in an axis of symmetry then they are equivalent genomes. These actions (rotation and reflection in an axis) generate the dihedral group D_n , which is a subgroup of the full group of rearrangements \mathcal{G} .

The equivalence defined by such rotations and reflections amounts to taking a quotient of the group arising from the fixed-frame genome space by the dihedral group. Two genomes being reachable from each other by such actions means that their corresponding group elements are in the same coset in this quotient. In other words, each equivalence class of genomes under three-dimensional rotations corresponds to a coset in \mathcal{G}/D_n . Put very bluntly, “genomes are cosets”. Each element $g \in \mathcal{G}$ lies in a coset $gD_n = \{gd \mid d \in D_n\}$, so that elements within a coset represent the same arrangement. Instead of mapping the genomes to group elements; to account for symmetry the genomes are instead mapped to *cosets* of \mathcal{G} under D_n .

In this light, a measure of distance between two genomes needs to be distance between the *cosets* corresponding to each genome. This is implemented for minimal distances in Egri-Nagy et al. [2014a]. The minimal distance is found by searching through all ordered pairs of coset representatives, finding the pair which minimises the minimal distance (n regions implies $4n^2$ calculations, but the transitivity of the Cayley graph reduces this to $2n$). Formally, for cosets g_1D_n and g_2D_n , the minimal distance is given by

$$d_{\min}(g_1D_n, g_2D_n) = \min\{d(h_1, h_2) \mid h_1 \in g_1D_n; h_2 \in g_2D_n\}.$$

From group transitivity, we need not check all pairs: it suffices to compare the minimal distances from the elements of one coset to an arbitrary fixed representative of the other coset.

3. LIKELIHOOD FUNCTIONS FOR GROUP ELEMENTS

We now begin the construction of likelihood functions for evolutionary distance. While genomes are regarded as cosets of group elements that are equivalent under the dihedral group action, we begin with considering group elements alone, before building the genome likelihood functions in Section 6.

An important point with regard to group-based genome rearrangement models is that the group is generally *non-abelian*, which means that operations do not commute i.e. the order in which they occur is of consequence. This arises directly from the biological model: the effect of two successive inversions that overlap depends on the order in which they are done. This is very different to single nucleotide changes, that are assumed to occur independently. The independence assumption for SNPs means that mutation events at different sites commute.

When it comes to random walks, the key impact is that with abelian groups all endpoints of a path of a given length are equally likely. This is not the case for non-abelian groups such as those generated by inversion models, as can be seen in the following example.

Example 3.1. *Consider the group \mathcal{G} generated by linear (as opposed to circular) 2-inversions over 9 regions so that $\mathcal{S} = \{s_i = (i, i+1) \mid i \in 1, \dots, 8\}$. Both group elements $g_1 = (1, 4, 3, 2)$ and $g_2 = (2, 3)(4, 5)(8, 9)$ have minimal distance of three. While g_1 can only be realised by the sequence $s_1 s_2 s_3$; g_2 can be realised by $s_2 s_4 s_8$, $s_4 s_2 s_8$, $s_8 s_4 s_2$, and more. This particular example relies on the fact that disjoint cycles commute.*

Let g be a genome arrangement, with n the number of rearrangement events allowed by the model. Write $\alpha_i(g)$ for the number of paths from e to g of length $i \in \mathbb{N}$. Parameterise $\lambda = rT$ where T is time and r the number of rearrangements per unit time. Then the likelihood of λ given the path ends at g is given by

$$\begin{aligned} L(\lambda \mid g) &= Pr(g \mid \lambda) \\ (1) \quad &= \sum_{i \geq 0} Pr(g \mid i) Pr(i \mid \lambda) \end{aligned}$$

Assuming that time between events follows an exponential distribution, we have $Pr(i \mid \lambda) = e^{-\lambda} / (\lambda^i i!)$. The assumption that paths of equal length are equally likely forces $Pr(g \mid i) = \alpha_i(g) / n^i$, and so

$$(2) \quad L(\lambda \mid g) = \sum_{i \geq 0} \frac{e^{-\lambda} \lambda^i}{i!} \frac{\alpha_i(g)}{n^i}.$$

Maximising this function with respect to λ gives a maximum likelihood estimate $\hat{\lambda}$ of this parameter. In some special cases, closed-form expressions for $\alpha_i(g)$ may yield closed form likelihood functions (such as in Example 3.2). In other cases, $\hat{\lambda}$ can be obtained numerically.

Example 3.2. *A circular genome with only three regions under a model of inversions of adjacent pairs of regions corresponds to the symmetric group S_3 with*

circular generators $\{(1, 2), (2, 3), (3, 1)\}$. One can show that $\alpha_i(g) = 3^{i-1}$ if g and i are even and odd, and if $d_{\min}(g) \geq i$ (and zero otherwise). In this simple example the likelihood functions are

$$\begin{aligned} L(\lambda \mid ()) &= \frac{e^{-\lambda}}{3} \left[3 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} \dots \right] = \frac{e^{-\lambda}}{3} [2 + \cosh \lambda]; \\ L(\lambda \mid (1, 2)) &= \frac{e^{-\lambda}}{3} \left[\frac{\lambda^1}{1!} + \frac{\lambda^3}{3!} + \frac{\lambda^5}{5!} \dots \right] = \frac{e^{-\lambda}}{3} \sinh \lambda; \\ L(\lambda \mid (1, 2, 3)) &= \frac{e^{-\lambda}}{3} \left[\frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} \dots \right] = \frac{e^{-\lambda}}{3} [\cosh \lambda - 1]. \end{aligned}$$

The likelihood functions in Example 3.2 are monotonic in λ ; something that is not true in general. For more realistic models with more regions, no closed form expressions are known, and hence the likelihood functions must be approximated by truncating the series.

Unlike cases such as Example 3.2, the path-count function $\alpha_i(g)$ does not usually have a closed form. It can, however, be computed using a simple recursive algorithm. Suppose we want to count the number of paths of length i that end at g . Each such path goes through an immediate neighbour of g , after having traversed a path of length $i-1$. Therefore the number of paths of length i to g is the sum of the numbers of paths of length $i-1$ to the immediate neighbours of g . While there are some economies that can be made to this recursion (for instance we may know that some of the path-counts are zero), it is still a computationally demanding algorithm, and only currently effective in practice for models of up to nine regions.

4. MINIMAL DISTANCE AND MLE

In general, phylogenetic distance methods assume some relationship between distances and evolutionary time. That is, all methods presume that a larger distance implies a greater time since evolutionary divergence. In situations that enjoy independence among sites (such as SNP models) this is reasonable; however the motivation for this work is that it may not necessarily also hold for models generated by large scale rearrangements (which are generally non-abelian).

One way to investigate this relationship is to compare the relative orderings placed by the metric on the set of pairs of genomes. While distance based phylogeny reconstruction methods do not rely solely on the ordering of distances, they are sensitive to it.

Figure 2 highlights examples where the partial order under the minimal distance and the partial order under the MLE differ (i.e. $d_{\min}(g_1) > d_{\min}(g_2)$ but $\hat{\lambda}_{g_1} < \hat{\lambda}_{g_2}$). This reversal of partial order relations is (given our assumptions) a function of path counts. These examples are not uncommon and highlight the problem with the minimal distance: it does little to characterise the MLE. It is not difficult to construct examples where this gives rise to differences in phylogenetic inference (see Figure 6). We turn our attention to the question of what conditions on arrangements give rise to the same MLE.

5. WHAT GROUP ELEMENTS HAVE THE SAME MLE?

The previous section demonstrated that the structure of the genome space, represented by the Cayley graph, makes minimal distance a poor proxy of evolutionary

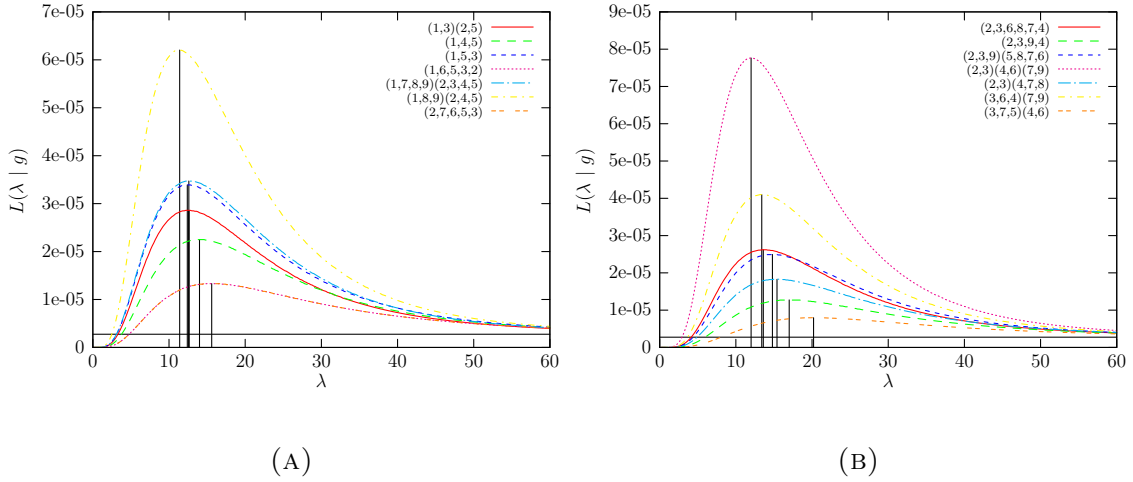


FIGURE 2. Samples of elements of minimal distance six (Fig 2a), and minimal distance seven (Fig 2b). The estimates, shown as vertical lines for each likelihood curve, clearly overlap and in some cases the partial order reversal is stark: $d_{\min}((2, 7, 6, 5, 3)) = 6$ and $d_{\min}((2, 3)(4, 6)(7, 9)) = 7$ but $\hat{\lambda}_{(2, 7, 6, 5, 3)} \approx 16$ and $\hat{\lambda}_{(2, 3)(4, 6)(7, 9)} \approx 12$.

time. A natural question arises: are there features of arrangements that can predict the MLE? One possibility for two arrangements to have the same MLE is when they have identical likelihood functions, and likelihood functions are determined by path counts (Eq. (2)). In Clark et al. [2016], it is shown that if two group elements are conjugate under the normalizer of the generating set, then their set of minimal length paths are not only the same size, but also order isomorphic. An extension of this provides a sufficient condition among arrangements to ensure equality of the MLE.

Definition 5.1 (Normalizer). *Let \mathcal{G} be a group, and X a subset of \mathcal{G} . The normalizer of X in \mathcal{G} is defined as*

$$N_{\mathcal{G}}(X) = \{g \in \mathcal{G} \mid g^{-1}Xg = X\}.$$

Proposition 5.1. *Let \mathcal{G} be a group generated by \mathcal{S} . Write \sim_N to mean conjugate under an element of $N_{\mathcal{G}}(\mathcal{S})$. For $g_1, g_2 \in \mathcal{G}$ we have*

$$g_1 \sim_N g_2 \implies L(\lambda \mid g_1) \equiv L(\lambda \mid g_2) \implies \hat{\lambda}_{g_1} = \hat{\lambda}_{g_2},$$

where λ_1 and λ_2 represent the MLE for g_1 and g_2 respectively.

Proof. From Equation (2), the likelihood functions are of the form $L(\lambda \mid g) = e^{-\lambda}P(\lambda \mid g)$ where $P(\lambda \mid g)$ is a polynomial in λ . Two polynomials are equal if and only if their respective coefficients are equal. Hence we have that

$$\alpha_i(g_1) = \alpha_i(g_2) \forall i \geq 0 \implies L(\lambda \mid g_1) \equiv L(\lambda \mid g_2) \implies \hat{\lambda}_{g_1} = \hat{\lambda}_{g_2}.$$

This reduces the proof to showing that $g_1 \sim_N g_2$ implies $\alpha_i(g_1) = \alpha_i(g_2)$, for all $i \geq 0$.

Let $g \sim_N h$, so that $h = \pi^{-1}g\pi$ for some $\pi \in N_{\mathcal{G}}(\mathcal{S})$. We will show that there is a bijection between the paths of length i to g and to h , which will show that $\alpha_i(g) = \alpha_i(h)$ for all i .

Let $R_i(g)$ be the set of all length i paths to g , so that $|R_i(g)| = \alpha_i(g)$. Take $\gamma_g \in R_i(g)$ to be a path realised as a concatenation of i generators;

$$\gamma = s_{k_1} s_{k_2} \dots s_{k_i}$$

so that

$$e\gamma = es_{k_1} s_{k_2} \dots s_{k_i} = g.$$

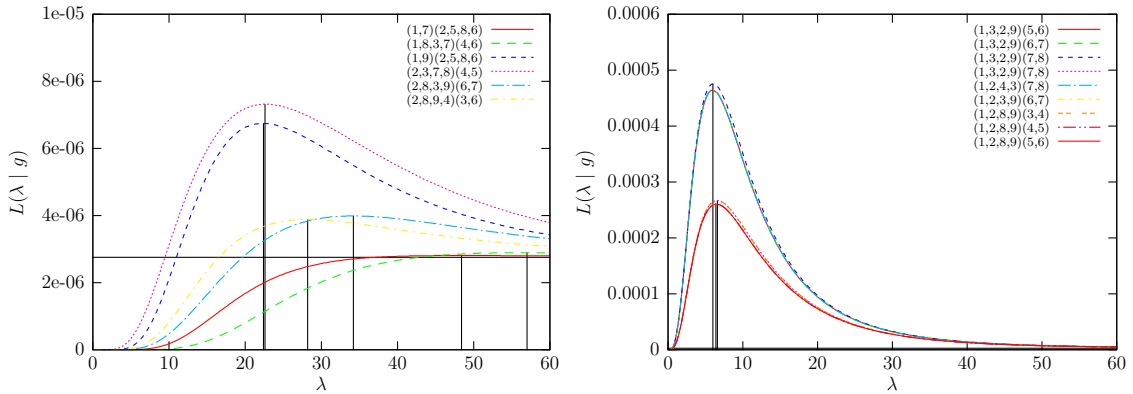
It follows that

$$\begin{aligned} h &= \pi^{-1} g \pi = \pi^{-1} e \pi \cdot \pi^{-1} s_{k_1} \pi \cdot \pi^{-1} s_{k_2} \pi \cdot \dots \cdot \pi^{-1} s_{k_i} \pi. \\ &= es_{l_1} s_{l_2} \dots s_{l_i} \end{aligned}$$

where $s_{l_n} = \pi^{-1} s_{k_n} \pi$. Because $\pi \in N_{\mathcal{G}}(\mathcal{S})$ and $s_k \in \mathcal{S}$, we have each $s_l \in \mathcal{S}$. Therefore, $s_{l_1} s_{l_2} \dots s_{l_i}$ defines a path to h of length i , and so conjugation by π is an injective map from $R_i(g) \rightarrow R_i(h)$.

For surjectivity, each path $\gamma_h \in R_i(h)$ has pre-image $\pi\gamma\pi^{-1} \in R_i(g)$. Hence $g_1 \sim_N g_2 \implies \alpha_i(g_1) = \alpha_i(g_2)$. \square

If elements are conjugate (but not by $N_{\mathcal{G}}(\mathcal{S})$) their likelihood functions are generally different (see Figure 3b) and are not guaranteed to give the same MLE. In Example 3.2, $N_{\mathcal{G}}(\mathcal{S})$ is the entire group, and in this case all elements of a conjugacy class will all also be conjugate under the normalizer. We now return to the calculation of likelihood functions for distances between two genomes, allowing movement in three dimensions.



(A) Group elements of various d_{min} from the same conjugacy class.

(B) Group elements identical minimal distance ($d_{min}(g) = 4$) in C .

FIGURE 3. Likelihood functions and MLEs from the conjugacy class C with representative $(1, 2, 3, 4)(5, 6)$.

6. MLE ESTIMATES OF DISTANCE BETWEEN GENOMES AS COSETS

Given the likelihood functions defined above for particular group elements we now introduce their application to cosets in order to address the biological issue of dihedral symmetry. For the dihedral group $D_n \subseteq G$, define the cosets

$$gD_n = \{gd \mid d \in D_n\}.$$

The cosets partition the group, each of size $2n$. A genome corresponding to a group element in turn corresponds to an element of one coset. Each element of a particular coset represents all orientations of a particular permutation, and hence a genome arrangement is now seen as a *coset* of \mathcal{G} under D_n .

To construct a likelihood function for pathlength between two cosets, in principle we need to consider all paths between the cosets. Fortunately, we can use the results of the previous section to reduce the need to count paths between these $(2n)^2$ pairs of elements.

Take two cosets g_1D_n and g_2D_n ($g_1, g_2 \in G$), and let γ be a path of length i between g_1 and g_2 , so that $g_1\gamma = g_2$. Then for any element of the dihedral group $d \in D_n$ we have $g_1\gamma d = g_2d$ and therefore

$$g_1d(d^{-1}\gamma d) = g_2d.$$

Note that $g_1d \in g_1D_n$ and $g_2d \in g_2D_n$, so for each $d \in D_n$, $d^{-1}\gamma d$ defines another path between the two cosets.

Since D_n is a subgroup of the normaliser of the generating set, $N_{\mathcal{G}}(\mathcal{S})$, the element $d^{-1}\gamma d$ remains a path of length i , by the arguments in the proof of Proposition 5.1. This means that to count paths of length i between the two cosets, it is sufficient to choose a single representative of one of the cosets and consider paths of length i to each of the $2n$ elements of the other coset. In other words, instead of considering paths of length i between g_1D_n and g_2D_n , we may simply consider paths between g_1 and g_2D_n .

The other simplification that can be made is that as before, the transitivity of the Cayley graph under left multiplication means that instead of paths from g_1 to g_2D_n we may instead count paths from the identity e to $g_1^{-1}g_2D_n$. The problem is reduced to considering paths of length i from the identity to any coset gD_n .

Let g be one permutation representation of the genome, and gD_n the corresponding coset. Because paths to each coset element are independent from each other, the likelihood function splits into a sum across elements of the coset:

$$\begin{aligned} L(\lambda \mid X = gD_n) &= Pr(e \rightarrow gD_n \mid \lambda) \\ &= \sum_{d \in D_n} Pr(gd \mid \lambda) \\ &= \sum_{d \in D_n} \sum_{i \geq 0} Pr(gd \mid i) \cdot Pr(i \mid \lambda) \\ &= \sum_{d \in D_n} \sum_{i \geq 0} \frac{\alpha_{gd}(i)}{n^i} \cdot \frac{e^{-\lambda} \lambda^i}{i!}. \end{aligned}$$

The last expression above has terms that are just the likelihoods for individual group elements, derived in Section 3, and so we have:

$$(3) \quad L(\lambda \mid X = gD_n) = \sum_{d \in D_n} L(\lambda \mid X = gd).$$

That is, the likelihood for a genome distance, allowing for three-dimensional rotations, is the sum of the individual likelihood functions for each of the group elements in the coset.

7. MINIMAL DISTANCES AND MLEs OF DISTANCES BETWEEN GENOMES.

We have now described several ways to define a distance between genomes under a group-theoretic model of rearrangement (such as using inversions). Given one reference genome and another given by a group element g (describing the permutation of the regions), these are:

- (1) The minimal distance to g ,
- (2) The minimal distance to the coset containing g (the method developed in Egri-Nagy et al. [2014b] to account for the genome in three-dimensional space),
- (3) A maximum likelihood estimate for the distance to g , and
- (4) A maximum likelihood estimate for the distance to the coset containing g .

The latter two have been introduced in this paper, and we have described the way the MLE for the distance to a group element (3) can provide more information than use of the minimal distance alone (1), in the sense that the ordering of elements by minimal distance is often not preserved when taking MLEs (Figure 2).

The same clash between minimal distance and MLE of distance arises when allowing the genome to rotate in three dimensions (the coset approach). For instance, consider the group element $(1, 2)$ with minimal distance 1 from the identity. If we rotate the arrangement on nine regions once, we obtain the group element $(1, 3, 4, 5, 6, 7, 8, 9)$ (region 1 is in position 3, region 3 is in position 4 etc), which has minimal distance 7. Another rotation gives $(1, 4, 6, 8)(2, 3, 5, 7, 9)$ with minimal distance 13. In other words, a single coset can contain elements of very different minimal lengths. This is a reminder that fixing the frame of reference and calculating the minimal distance between two genomes on the basis of just one frame is likely to result in large errors.

The occasional partial order reversal (with respect to minimal distance) observed for group elements persists for the coset case. This can be seen in Figures 4a and 4b, in which MLEs for cosets whose minimal distances are 6 and 7 are shown: some cosets with minimal distance 6 have higher MLE than some cosets with minimal distance 7. This confirms that minimal distance, even when used on cosets, may be a poor tool for estimating pairwise distance.

The existence of an MLE for an arrangement is not guaranteed, whether one considers it fixed in space (as a single group element) or free to rotate (as a coset). Indeed, even if a particular group element gives an MLE, there is no guarantee the coset it resides in also gives an MLE. Similarly, many cosets with MLEs will contain elements that do not individually possess an MLE. Figure 5 shows examples of the likelihood functions for elements within the same coset. The two cosets, $(1, 2, 7)D_9$ and $(1, 8, 2, 9, 3, 4)D_9$, contain 6 and 5 group elements respectively (out of a total of 18 elements), that individually possess MLEs. Here, not just the value of the MLE, but also the probability associated with each MLE is important. For a coset to have an MLE, the MLEs of the group elements it contains must be “strong enough” to persist through the construction of the coset likelihood function. This illustrates a previous point – many group elements which themselves possess no MLE, may reside in a coset which *does*.

For a genome with n regions, each coset contains $2n$ elements. The likelihood function for cosets, as in Equation 3, is a sum of the likelihood functions of the elements in the coset. By accounting for dihedral symmetry, the size of the space

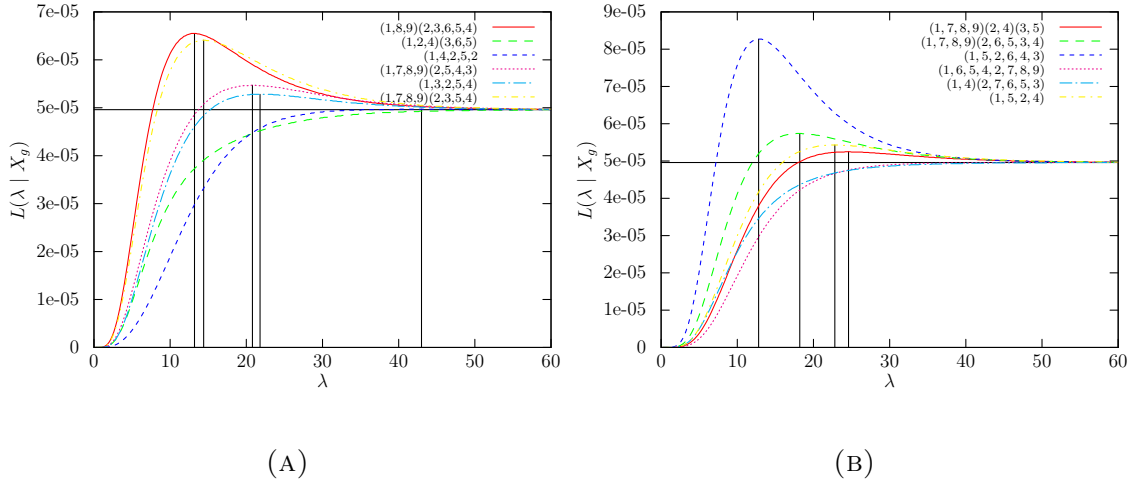


FIGURE 4. Likelihood functions and MLEs where the genome arrangement is viewed as a coset of \mathcal{G} under D_n . Figures 4a and 4b show cosets of minimal distance 6 and 7 respectively. The cosets gD_n are labelled by a single representative element.

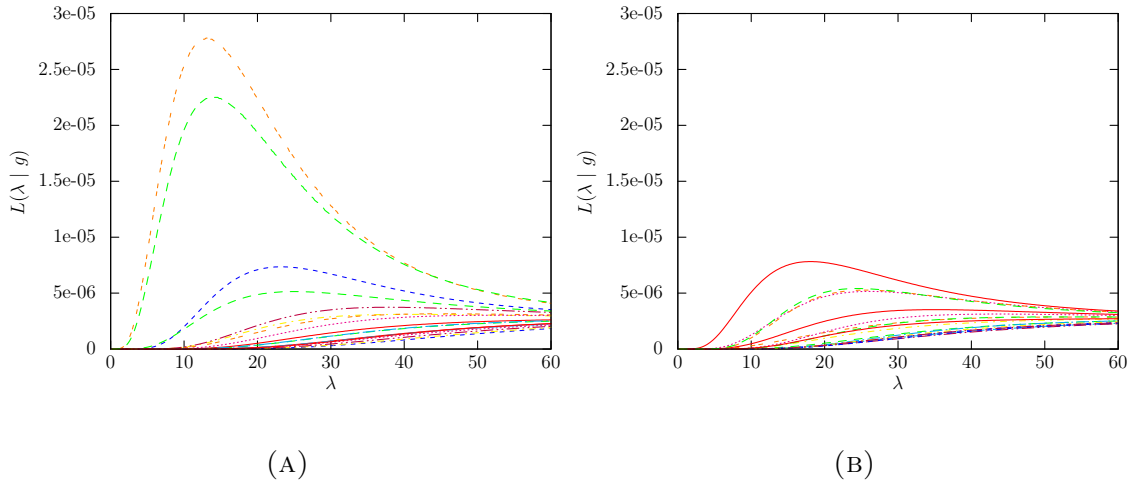


FIGURE 5. Likelihood functions from within two cosets. Figure 5a (resp. 5b) shows elements of the coset represented by (1,2,7) (resp. (1,8,2,9,3,4)).

reduces by a factor of $2n$: S_n has order $n!$, while S_n/D_n has order $n!/2n$. An exhaustive calculation of $\mathcal{G} = S_9$ shows $\sim 41\%$ of group elements possess an MLE. This represents a lower bound on this proportion, as firstly, all terms of the likelihood function are positive (and so a detected maximum will stay a maximum), and secondly it is possible that new maxima may be found beyond the truncation. By comparison, treating genomes as cosets reveals a lower bound of $\sim 44\%$.

8. THE EFFECT OF THE USE OF THE COSET MLE ON PHYLOGENY

We have seen that the use of a maximum likelihood estimator for evolutionary distance can change the ordering on genome distances. One would expect this to have a significant effect on phylogenetic inference, and it does. We show in

Figure 6 an example of phylogenies obtained on four genomes, each containing the same nine regions. These phylogenies are obtained using the neighbour joining algorithm [Saitou and Nei, 1987], based on distances obtained from the four methods listed in Section 7.

The phylogenies on the left of Figure 6 are obtained using the minimal distance methods, and those on the right are obtained using maximum likelihood estimates. The phylogenies on the top are obtained from distances between single group elements, and those on the bottom from distances between cosets.

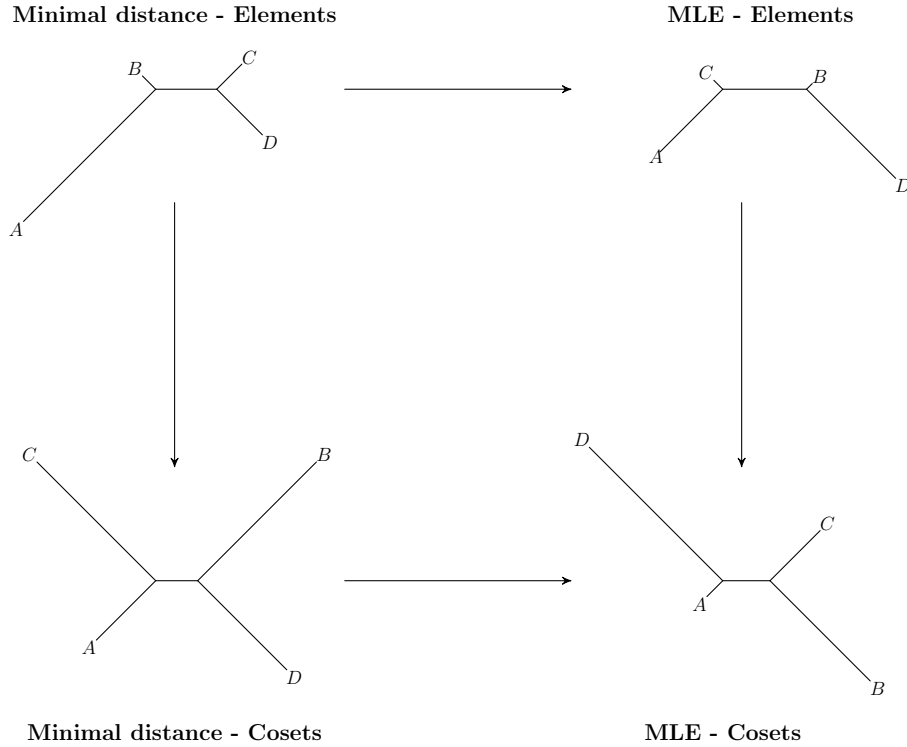


FIGURE 6. Four phylogenies obtained using different distance measures on the same four genomes: $A = ()$, $B = (1, 3, 2, 5, 4)$, $C = (1, 7, 8, 9)(2, 5, 4, 3)$, and $D = (1, 5, 2, 6, 4, 3)$. The top left phylogeny uses the fixed frame minimal distance; top right results from using the MLE approach on a fixed frame. The bottom left represents the coset minimal distance unfixing the frame of reference, and the bottom right results from the MLE approach on cosets. Note that the edge length in the top right diagram between the element $(1, 3, 2, 5, 4)$ and the bifurcation was negative, an occasional issue with Neighbour joining.

The four quartet phylogenies shown in Figure 6 differ significantly in their topologies. We see $AB|CD$ in the top left (minimal distance on a single group element); $AC|BD$ in the top right and bottom left (MLE on a single group element and minimal distance on a coset); and $AD|BC$ in the bottom right hand corner, the preferred MLE on the coset.

9. DISCUSSION

This paper introduces a maximum likelihood estimator for the evolutionary distance between two genomes under a large-scale genome rearrangements model. One may view this as a correction method for a family of models which can be interpreted using non-abelian groups. Methods of correcting distances for multiple changes are commonly used, because the use of uncorrected distances can lead to poor inference regarding topology (see [Felsenstein \[2004\]](#)). These corrections for multiple changes are typically implemented in the context of single nucleotide polymorphisms, and these are typically in an environment in which changes at each site are considered to be independent. The large-scale rearrangements discussed in this paper are different from this in several ways, but the key difference is that rearrangements can affect overlapping regions, and hence interact with each other (giving rise to a non-abelian group model). The “correction” involved in the context of this paper is to account for evolutionary paths between two genomes that might not be the shortest path.

A feature of this approach is that we have addressed the problem of fixing a frame of reference by taking into account the action of the dihedral group. This means that we consider arrangements to be equivalent if they can be obtained by physically rotating the genome in three dimensional space. Algebraically, this involves taking a quotient by the dihedral subgroup and treating each genome as a coset under this quotient.

One may view the maximum likelihood estimator in this paper, then, as a *two-fold* correction for any set of rearrangements which generate a group. That is, for models described by the formalism in [Egri-Nagy et al. \[2014b\]](#), including any combination of inversions, translocations, and adjacent transpositions. The results can be applied to any group/generating set, and hence can also be used in signed permutations groups to account for the relative orientation of regions.

The MLE approach that we have described requires a selection of generators corresponding to legal rearrangements, and an assumption regarding the probability distribution across these generators. In our examples, we have used a uniform distribution over the set of circular adjacent transpositions for simplicity. As noted already, the generating set can be readily changed, however it is also possible to change the probability distribution on these generators. For instance, a natural example might be to model both inversions and translocations, and assign different probabilities to each type of event. When a non-uniform distribution such as this is selected, the $P(g | i)$ factor in Equation 1 is no longer just a fraction based on path counts, but can nevertheless still be calculated using similar methods (and with the same complexity).

The likelihood functions of group elements, and of cosets, do not always have maxima, and in this situation we are not able to give an estimate of evolutionary distance. This is analogous to the limits of the Jukes-Cantor correction for SNP models of evolution, which are also unable to give a distance when the proportion of nucleotides that have changed exceeds 0.75. An interesting open question is to characterise the genomes (or group elements) for which the MLE exists. As described in previous sections, one experiment on genomes of 9 regions under a model of adjacent inversions found slightly more than half of the genomes have MLEs. Since taxa under study are likely to be in a reasonably contained region of the full genome space, this suggests that in most empirical studies MLEs will exist.

The key challenges to the MLE approach described here are computational. The likelihood function for a particular element g consists of two main probabilities: $Pr(i \mid \lambda)$ and $Pr(g \mid i)$. While the first is trivial under our assumptions, the second is computed as the proportion of paths of length i which end at g . The current path count algorithm relies on dynamic programming with complexity exponential in $|\mathcal{S}|$ and memory factorial in n . A compromise must be found between truncation and accuracy of the MLE before this can be applied to genomes of realistic numbers of regions (say, greater than 30). The structure provided by the group model gives some hope; Lemma 5.1 is an example of an algebraic property that can be used to significantly reduce computation time. For truncated lookup table construction, this property alone results in a $2n$ -fold decrease in computation time.

The MLE approach introduced here represents a new distance measure between two arrangements. The MLE of true evolutionary distance provides a measure which accounts for underlying path structure and by construction addresses the selection of frame of reference in a natural manner. Possible future work is abundant; centered around computational tractability and the ability to generalise to different generating sets and probability measures.

REFERENCES

- David Aldous and Jim Fill. Reversible Markov chains and random walks on graphs. preprint available at <http://stat-www.berkeley.edu/users/aldous/book.html>., 2002. 3, 4
- Kevin Atteson. The performance of Neighbor-Joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2-3):251–278, 1999. 2
- László Babai. Automorphism groups, Isomorphism, Reconstruction. In *Handbook of Combinatorics (vol. 2)*, pages 1447–1540. MIT Press, 1996. 4
- Peter Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences*, 1971. 2
- Peter Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50, 1974. 2
- Chad Clark, Attila Egri-Nagy, Andrew R Francis, and Volker Gebhardt. Bacterial phylogeny in the Cayley graph. *arXiv preprint arXiv:1601.04398*, 2016. 3, 8
- Daniel Dalevi and Niklas Eriksen. Expected gene-order distances and model selection in bacteria. *Bioinformatics*, 24(11):1332–1338, 2008. 2
- Richard Durrett, Rasmus Nielsen, and Thomas L York. Bayesian estimation of genomic distance. *Genetics*, 166(1):621–629, 2004. 2
- Attila Egri-Nagy, Andrew R Francis, and Volker Gebhardt. Bacterial genomics and computational group theory: The biogap package for gap. In *Mathematical Software–ICMS 2014*, pages 67–74. Springer, 2014a. 5
- Attila Egri-Nagy, Volker Gebhardt, Mark M Tanaka, and Andrew R Francis. Group-theoretic models of the inversion process in bacterial genomes. *Journal of Mathematical Biology*, 69(1):243–265, 2014b. 3, 4, 11, 14
- Niklas Eriksen. Approximating the expected number of inversions given the number of breakpoints. In *Algorithms in Bioinformatics*, pages 316–330. Springer, 2002. 2
- Niklas Eriksen. Expected number of inversions after a sequence of random adjacent transpositionsan exact expression. *Discrete Mathematics*, 298(1):155–168, 2005. 2

- Niklas Eriksen and Axel Hultman. Estimating the expected reversal distance after a fixed number of reversals. *Advances in Applied Mathematics*, 32(3):439–453, 2004. [2](#)
- Joseph Felsenstein. *Inferring Phylogenies*. Sinauer associates Sunderland, 2004. [1](#), [14](#)
- Guillaume Fertin. *Combinatorics of genome rearrangements*. MIT press, 2009. [1](#)
- Andrew R Francis. An algebraic view of bacterial genome evolution. *Journal of Mathematical Biology*, 69(6-7):1693–1718, 2014. [3](#)
- O. Gascuel, editor. *Mathematics of Evolution and Phylogeny*. OUP Oxford, 2005. [1](#)
- Olivier Gascuel and Mike Steel. *Algorithmica*, pages 1–18, 2015. [2](#)
- Chris Godsil and Gordon Royle. Algebraic Graph Theory, volume 207 of Graduate Texts in Mathematics, 2001. [4](#)
- Thomas H Jukes and Charles R Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132, 1969. [2](#)
- Wang Li-San. Genome rearrangement phylogeny using weighbor. In *Algorithms in Bioinformatics*, pages 112–125. Springer, 2002. [2](#)
- Alexander Lubotzky. Cayley graphs: Eigenvalues, expanders and random walks. *London Mathematical Society Lecture Note Series*, pages 155–190, 1995. [4](#)
- Roger C Lyndon and Paul E Schupp. *Combinatorial Group Theory*. Springer, 1977. [4](#)
- István Miklós. MCMC genome rearrangement. *Bioinformatics*, 19(suppl 2):ii130–ii137, 2003. [2](#)
- István Miklós and Aaron E Darling. Efficient sampling of parsimonious inversion histories with application to genome rearrangement in yersinia. *Genome biology and evolution*, 1:153–164, 2009. [2](#)
- István Miklós and Eric Tannier. Bayesian sampling of genomic rearrangement scenarios via double cut and join. *Bioinformatics*, 26(24):3012–3019, 2010.
- Bernard ME Moret, Li-San Wang, Tandy Warnow, and Stacia K Wyman. New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17(suppl 1):S165–S173, 2001. [2](#)
- Joseph H Nadeau and Benjamin A Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences*, 81(3):814–818, 1984. [2](#)
- Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987. [2](#), [13](#)
- Jeremy G Sumner, Jesús Fernández-Sánchez, and PD Jarvis. Lie markov models. *Journal of theoretical biology*, 298:16–31, 2012.
- Li-San Wang and Tandy Warnow. Estimating true evolutionary distances between genomes. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 637–646. ACM, 2001. [2](#)
- Tandy Warnow. Some combinatorial problems in phylogenetics. In *Proc. Int’l Colloquium on Combinatorics and Graph Theory*, 1996. [2](#)
- Thomas L York, Richard Durrett, and Rasmus Nielsen. Bayesian estimation of the number of inversions in the history of two chromosomes. *Journal of Computational Biology*, 9(6):805–818, 2002. [2](#)